Predikce objemu S&P 500

Michal Najman, najmanm@gmail.com

Tato zpráva pojednává o mnou navrženém prediktivním modelu střední hodnoty objemu denních obchodů v rámci indexu S&P 500. Finální model je sestaven z afinní kombinace obecně nelineárních složek zvolených na základě teoretických a empirických vlastností distribuce dat. Při porovnání s referenčním modelem dosahuje sestavený model výrazně lepších výsledků v rámci kritéria R2.

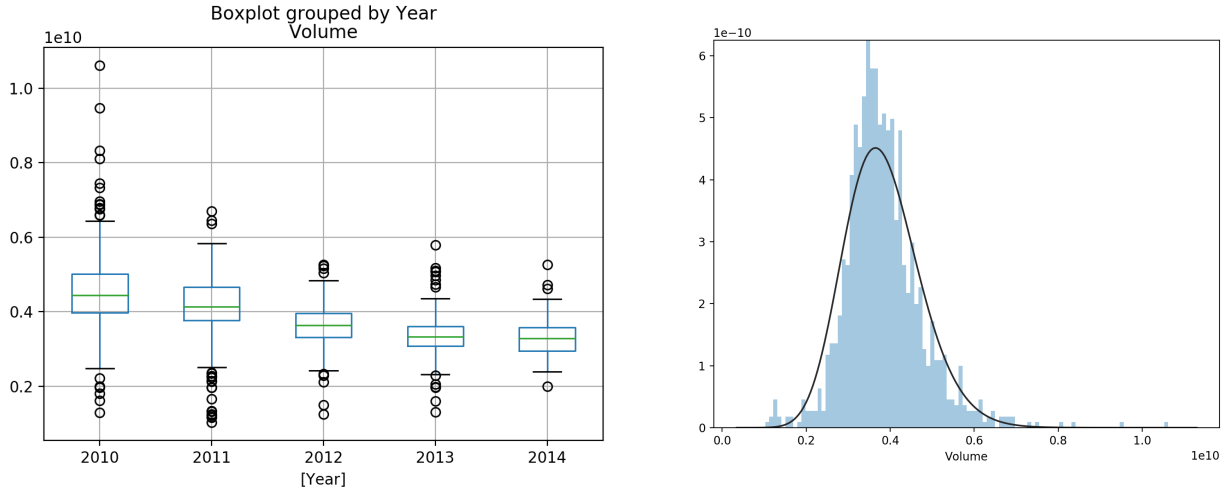
I. Dataset

Data z S&P 500 obsahují mimo jiné hodnotu denního objemu obchodů, kterou reprezentuji náhodnou veličinou V_d. Celkem se ve vymezeném období 1.1.2010-31.07.2014 obchodovalo 1151 dní. Byť bylo v zadání doporučeno uvažovat i jiné vstupní zdroje, k tomuto kroku jsem se nedostal. Mezi takové vstupy by například mohlo patřit: změna položek v indexu nebo významné zprávy o hospodaření. Abych předešel nejasnostem, v textu místo veškeré informace $\mathcal{F}_d$, která je uvedena v zadání úkolu, používám pouze historii hodnot objemu do dne d (včetně) $H_d = (V_d, V_{d-1}, \dots)$.

Časová řada přirozeně poskytuje pro každou proměnnou V_d jednu realizaci v_d, realizaci historie značím $h_d = (v_d, v_{d-1}, \dots)$, realizaci celé časové řady pak $h := h_{1151}$. Hodnoty proměnných jsou uváděny v dolarech, nicméně označení měny je dále vypuštěno.

Nelze předpokládat, že všechny V_d mají stejnou distribuci. Zejména distribuce proměnných, které jsou časově vzdálené, se liší. Při porovnání statistiky h se v závislosti na roku mění medián i kvantily.

Co se týče konkrétní podoby distribuce V_d, předpokládám, že se jedná o jednu z „bell-shaped„ distribucí – jelikož má vhodné vlastnosti, zvolil jsem lognormální rozdělení. Porovnání histogramu a hustoty lognormálního rozdělení realizace h ukazuje, že tato volba není přesná, nicméně poměrně dobře empirické rozdělení aproximuje. Faktorem nepřesnosti je mj. i výše zmíněný fakt, že proměnné v_d jsou realizace V_d, a tedy nejsou i.i.d. (dokonce ani nezávislé ani nemají stejné rozdělení).



Vlevo: Znázornění statistiky realizace h podle roku. Lze vidět, že parametry distribuce se mění: posouvá se medián a rozptyl; extrémní hodnoty se přibližují ke středu. Vpravo: Histogram realizace h a lognormální distribuce kalibrovaná h metodou maximální věrohodnosti.

II. Model

Při sestavování modelu jsem postupoval iterativně a zde prezentuji jen finální model, který dosáhl nejlepších hodnot R2. Více o tom, co jsem vyzkoušel, je rozepsáno v kapitole „Co jsem zkusil, ale nefungovalo to".

Finální model uvažuje závislost veličiny V_{d+1} na předchozích hodnotách. V potaz se bere jak průměr předchozích hodnot, tak medián, exponenciální průměr nebo aditivní „bonusy" za den v týdnu či extrémní předchozí hodnotu. Tyto obecně nelineární složky jsou spojeny afinní kombinací. Dále budu pracovat s modelem, kdy je náhodná veličina V_{d+1} dána vztahem:

(1) $$V_{d+1} = \beta_0 + \beta_1 \cdot \mathrm{LMA}(H_d) + \beta_2 \cdot \mathrm{EMA}(H_d) + \beta_3 \cdot \mathrm{MED}(V_d) + \mathrm{AQ}(V_d) + \mathrm{AWD}(d+1) + \epsilon_d$$

β_i	jsou koeficienty modelu
$\mathrm{LMA}(H_d)$	je linear moving average historie H_d
$\mathrm{EMA}(H_d)$	je exponential moving average historie H_d
$\mathrm{MED}(V_d)$	je medián V_d
$\mathrm{AQ}(V_d)$	je aditivní složka závislá na kvantilu V_d, tedy $\mathrm{AQ}(V_d) \sim q(V_d)$
$\mathrm{AWD}(d+1)$	je aditivní složka závislá na dni v týdnu, na který den $d+1$ připadá
ϵ_d	je nějaké náhodná veličina s nulovou střední hodnotu (šum)

II.A. Vztah pro střední hodnotu objemu

Jelikož úkolem je predikovat střední hodnotu $E[V_{d+1} \mid H_d]$, forma modelu se změní. Jednotlivé složky mají následující tvar.

(2) $$E[\mathrm{LMA}(H_d)] = \sum_{i=0}^{K} a_i \cdot E[V_{d-i}]$$

(3) $$E[\mathrm{EMA}(H_d)] = \frac{\sum_{i=0}^{K} \gamma^i \cdot E[V_{d-1}]}{\sum_{i=0}^{K} \gamma^i}$$

Pokud uvažujeme, že $V_d, V_{d-1}, \ldots, V_{d-(K-1)}$ jsou lognormální a mají přibližně podobné parametry rozdělení, pak pro střední hodnotu mediánu V_d platí:

(4) $$E[\mathrm{MED}(V_d)] \approx \left(\prod_{i=0}^{K-1} V_{d-i} \right)^{\frac{1}{K}}$$

(5) $$E[\mathrm{AQ}(V_d)] = E[X \mid q(V_d)]$$

(6) $$E[\mathrm{AWD}(d+1)] = E[X \mid d+1]$$

II.B. Odhad střední hodnoty objemu

Jelikož je k dispozici „pouze" jedna realizace v_d pro každou V_d, používám point-estimate střední hodnotyV_d, tedy $\tilde{E}[V_d \mid H_d] \approx v_d$. Model se tím redukuje na následující tvar:

(7) $$\tilde{E}[V_{d+1}\,|\,H_d] = \alpha_0 + \sum_{i=0}^{K_1} \alpha_1^i \cdot v_{d-i} + \alpha_2 \cdot \mathsf{MED}^{K_2}(h_d) + \alpha_3 \cdot \mathsf{EMA}^{K_3}(h_d) + \mathsf{AQ}^{K_4}(v_d) + \mathsf{AWD}(d+1)$$

Koeficienty α_m^n jsou parametry modelu a hodnoty K_m jsou hyperparametry jednotlivých složek modelu. V případě K_1, K_2, K_3 se jedná o časové horizonty, K_4 je vektor mezních hodnot dělení intervalu $\langle 0,1 \rangle$, viz rovnice (8).

Aditivní složka AQ je závislá na kvantilu. Jedná se o funkci, která vrací konkrétní číslo – v modelu lze toto číslo interpretovat jako odhad střední hodnoty korekce v závislosti na tom, v jakém intervalu se nachází kvantil $q(v_d)$. Tedy jedná se o korekci v případech, kdy je hodnota v_d extrémní.

(8) $$\mathsf{AQ}^{K_4}(v_d) = \begin{cases} \alpha_4^0, & q(v_D) < K_4^0 \\ \alpha_4^1, & K_4^0 \le q(v_D) < K_4^1 \\ \vdots & \\ \alpha_4^N, & K_4^{N-1} \le q(v_D) \end{cases}$$

Podobně, aditivní složka AWD je závislá na dni v týdnu. Vrací odhad střední hodnoty korekce pro pondělí, resp. ostatní obchodní dny.

(9) $$\mathsf{AWD}(d+1) = \begin{cases} \alpha_5^0, & d+1 \text{ je pondělí} \\ \vdots & \\ \alpha_5^4, & d+1 \text{ je pátek} \end{cases}$$

II.C. Mapa příznaků

Model tedy obsahuje následující složky:

1. offset;
2. lineární kombinaci realizací v_d;
3. odhad mediánu lognormálního rozdělení;
4. exponenciálně vážený průměr realizací v_d;
5. aditivní korekci podle toho, do kterého kvantilu v_d patří;
6. aditivní korekci podle dne v týdnu.

Prakticky však model reprezentuji pomocí příznakové mapy $\Phi(h_d)$, která dostupnou historii h_d zobrazí podle definice modelu na čísla v $\mathbb{R}$, resp. na $\{0,1\}$ u binárních složek. Pro číselné složky vrací $\Phi_m^n(h_d)$ odpovídající číselnou hodnotu a u aditivních korekcí je $\Phi_m^n(h_d)$ rovna booleanovské proměnné značící, jaký případ dle definice výše nastal.

Model tak lze zapsat ve tvaru vhodnějším pro učení parametrů.

(10) $$f(h_d) := \alpha_0 + \sum_{m,n} \alpha_m^n \cdot \Phi_m^n(h_d)$$

III. Učení modelu

Modely jsem vyhodnocoval dvěma způsoby: K-Fold učení a postupné učení v čase (rolling window evalation). K učení bylo použito několik algoritmů (k-NN, lineární regrese minimalizací SSE, neuronové

sítě a ensamblové metody), nakonec jsem pro finální model použil metodu epsilon-SVR (Support Vector Regression).

III.A. K-Fold Evaluation

Žádná použitá trénovací metoda nevyužívá fakt, že data jsou časová řada. Například nedělá tak explicitně ukládáním trénovacích vzorků a následným hledáním podposloupností podobných testovacímu příkladu. Z toho důvodu lze dataset pro daný účel považovat za i.i.d. (byť to je potenciálně nebezpečná relaxace) a model učit metodou k-fold s náhodným zamícháním. Aby testovacích dat bylo dostatek, a tedy byl snížen vliv extrémních náhodných hodnot v časové řadě, zvolil jsem $k = 5$, což odpovídá 20 % testovacích dat z původní množiny. Tedy přibližně 900 trénovacích a 230 testovacích vzorků.

III.B. Rolling Window Evaluation

Tato metoda více reflektuje reálnou aplikaci modelu a bere v potaz časovou závislost dat. Zvolením délky oken $\Delta_{TR} > 0$ a $\Delta_{TS} > 0$ se nastaví množství trénovacích a testovacích dat. Algoritmus pak iterativně pro všechna přípustná d natrénuje model na datech ze dní $\{d-\Delta_{TR},\ d-\Delta_{TR}+1,\ \ldots,\ d-1\}$ a následně otestuje na dnech $\{d,\ d+1,\ \ldots,\ d+\Delta_{TS}-1\}$. Tím vznikne graf závislosti evaluační funkce na d, který ukazuje které časové úseky jsou pro model „obtížnější".

Po vyzkoušení několika oken jsem nakonec používal stejné hodnoty $\Delta_{TR} = \Delta_{TS} = 300$. Tedy oproti K-Fold je výrazně méně trénovacích data, zato více testovacích.

Naměřené přesnosti u těchto dvou typů evaluací budou jiné. Vyšší přesnost v settingu K-Fold je způsobena náhodným zamícháním a jiným poměrem velikosti trénovací a testovací množiny. Pravděpodobnost, že se při testování objeví více extrémů, je nižší. Naopak u Rolling Window se extrémy projevují výrazně, dokud jsou obsaženy ve zvoleném testovacím okně, jelikož je respektována časová závislost řady. Proto budou přesnost pro identický typ modelu u Rolling Window nižší.

III.C. Normalizace

Metoda SVR je náchylná na rozsah vstupních data. Z toho důvodu po transformaci dat zobrazením $\Phi(h_d)$ jednotlivé složky normalizuji tak, aby trénovací data měla nulovou empirickou střední hodnotu a jednotkový empirický rozptyl. Tuto transformaci provádím jak pro příznakový vektor $\Phi(h_d)$, tak pro cílovou proměnnou v_{d+1}. Výstup SVR je po predikci transformován zpět do původního prostoru. Tento postup výrazně zlepšil výsledky modelu.

III.D. Naučené parametry

SVR při učení na jednotlivých iteracích nacházelo samozřejmě jinou sadu parametrů v závislosti na složení trénovací množiny, nicméně ty se příliš nelišily. Změna byla v řádu jednotek procent. V tabulce níže uvádím příklad naučených koeficientů modelu v jedné z iterací K-Fold (konkrétně pro seed 42). Parametry jsou zadány podle zápisu modelu (7), resp. (10).

Parametry modelu

Označení parametru	Odpovídají příznak	Hodnota parametru
α_0	offset	$43{,}23 \cdot 10^6$
α_1^0	v_d	$0{,}45$
α_1^1	v_{d-1}	$0{,}09$
α_2	MED	$0{,}24$
α_3	EMA	$0{,}18$
α_4^0	$\langle 0; 0{,}2)$	$124{,}30 \cdot 10^6$
α_4^1	$\langle 0{,}2; 0{,}8)$	$67{,}81 \cdot 10^6$
α_4^2	$\langle 0{,}8; 1\rangle$	$-231{,}79 \cdot 10^6$
α_5^0	pondělí	$-331{,}63 \cdot 10^6$
α_5^1	úterý	$152{,}09 \cdot 10^6$
α_5^2	středa	$92{,}34 \cdot 10^6$
α_5^3	čtvrtek	$75{,}79 \cdot 10^6$
α_5^4	pátek	$-18{,}28 \cdot 10^6$

Tabulka: Parametry modelu $f(h_d)$ naučeného metodou SVR. Pro porovnání, empirický průměr datasetu je $3\,839{,}84 \cdot 10^6$ a empirická standardní odchylka je $927{,}84 \cdot 10^6$. Lze tedy vidět, že aditivní korekce jsou opravdu jen korekce, které mírné zlepšují přesnost modelu – největší korekcí je penalizace za pondělí a zohlednění předchozí extrémně nízké hodnoty. Naopak významnou roli hraje dle koeficientů hodnota objemu obchodů z předchozího dne.

Nakonec jsem pro stanovení hyperparametrů nepoužil žádnou z kombinatorických metod, ale jen empiricky vyzkoušel, které hodnoty mají nejvyšší R2 v Rolling Window, popř. K-Fold evaluaci. V následující tabulce jsou uvedeny nalezené hodnoty.

Hyperparametry modelu a učení

Časový horizont v LMA, K_1	2
Časový horizont v MED, K_2	30
Časový horizont v EMA, K_3	7
Diskontní faktor v EMA, γ	0,9
Intervaly v QA, K_4	$\langle 0; 0{,}2), \langle 0{,}2; 0{,}8), \langle 0{,}8; 1\rangle$
Parametr C v SVR	1
Paremetr ϵ v SVR	0,05
Kernel SVR	lineární, tedy základní SVR

Tabulka: Hyperparametry finálního modelu.

III.E. Přesnost modelu

Model $f(h_d)$ je vyhodnocen kritériem R2 a MSE. Porovnání s referenčním modelem $\tilde{E}[V_{d+1} \mid H_d] = v_d$ je ukázáno v tabulce níže. Lze vidět, že naučený model $f(h_d)$ dosahuje vyšší přesnosti než referenční model. Podrobnější rozbor výsledků lze nalézt v další sekci.

K-Fold Evaluace

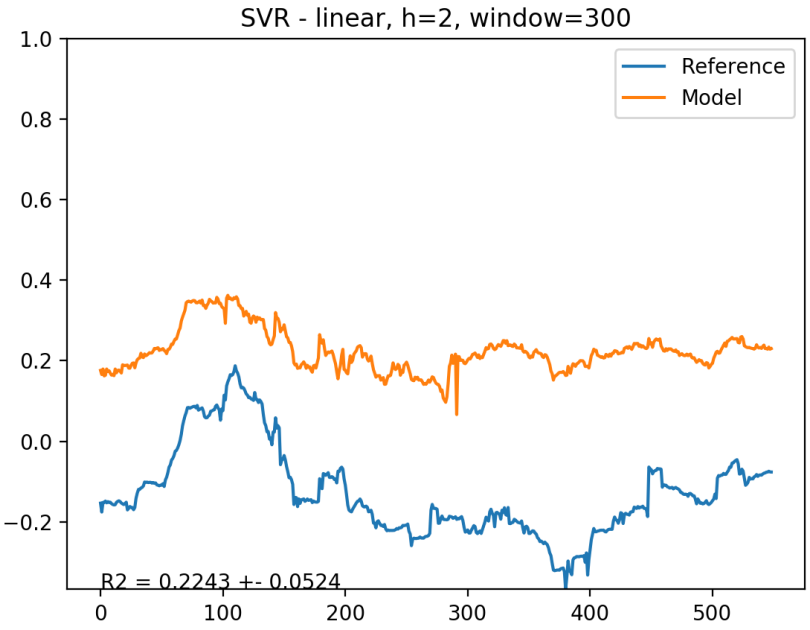
	MSE	R2	Průměrná abs. odchylka
Navržený model	$(625{,}44 \pm 42{,}56) \cdot 10^6$	$0{,}54 \pm 0{,}06$	$11{,}13 \pm 0{,}01\ \%$
Referenční model	$(731{,}74 \pm 55{,}20) \cdot 10^6$	$0{,}35 \pm 0{,}13$	$13{,}30 \pm 0{,}01\ \%$

Tabulka: Porovnání výsledků navrženého modelu a referenčního modelu v rámci K-Fold evaluace.

Rolling Window Evaluace

	MSE	R2	Průměrná abs. odchylka
Navržený model	$(524{,}34 \pm 61{,}07) \cdot 10^6$	$0{,}22 \pm 0{,}05$	$10{,}31 \pm 0{,}01\ \%$
Referenční model	$(630{,}53 \pm 65{,}09) \cdot 10^6$	-0.13 ± 0.11	$12{,}46 \pm 0{,}01\ \%$

Tabulka: Porovnání výsledků navrženého modelu a referenčního modelu v rámci Rolling Window evaluace.

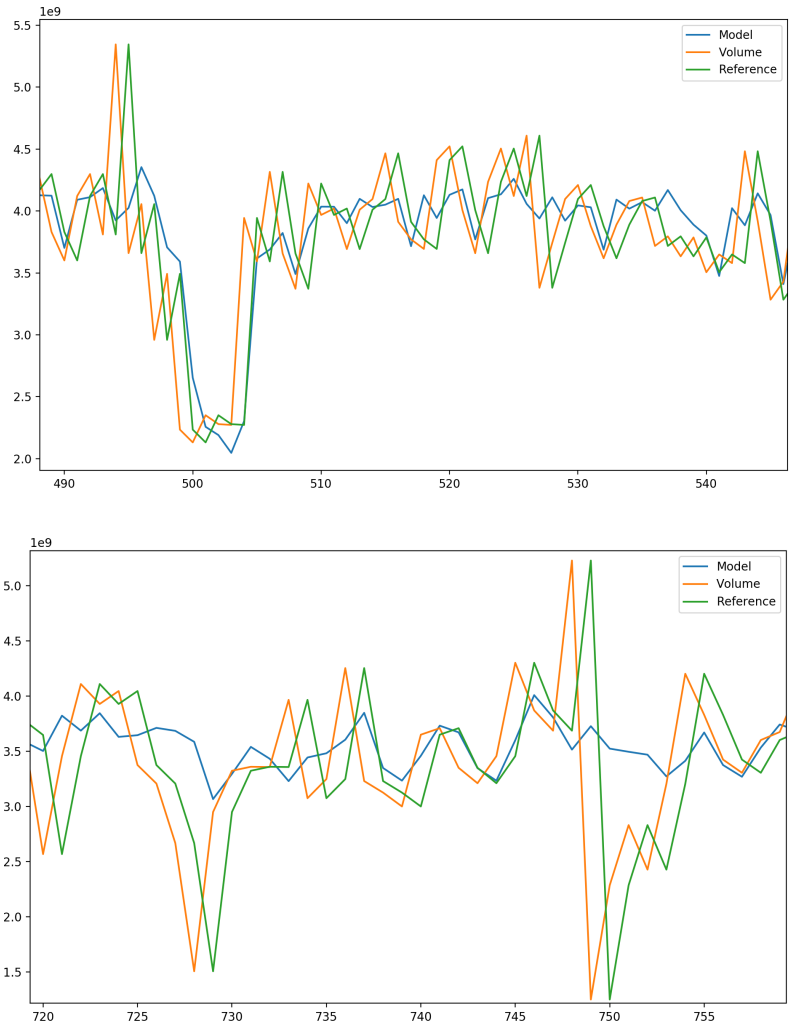


Obrázek: Závislost R2 na dni d v rámci Rolling Window vyhodnocení. Ze začátku časové řady lze vidět, že se přesnost referenčního modelu blíží navrženému modelu. Nicméně v ostatních oblastech je navržený model značně přesnější.

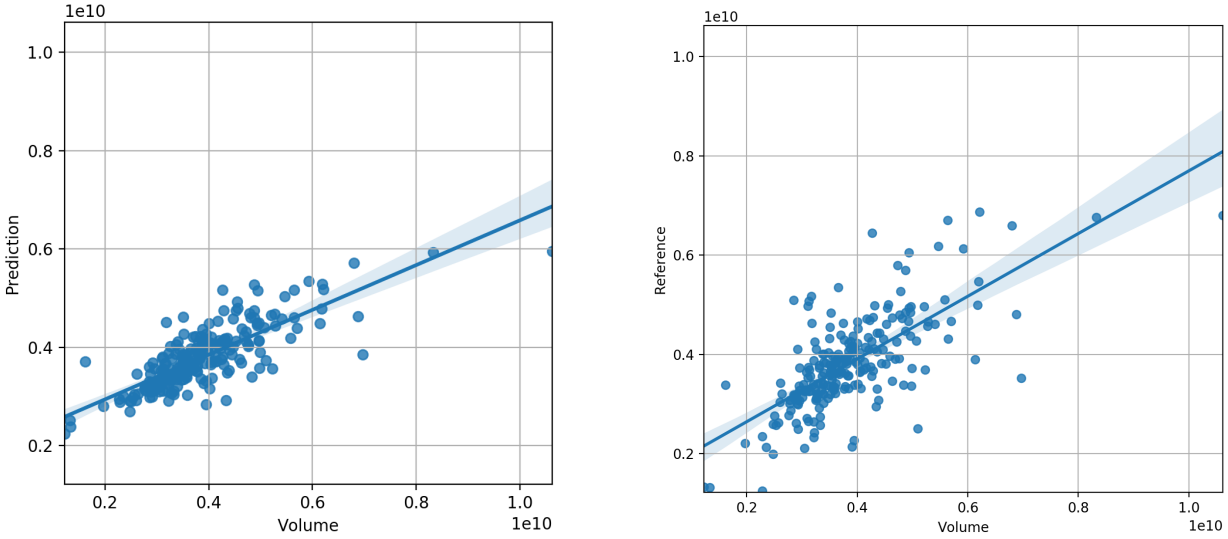
IV. Analýza výsledků

Navržený model celkem uspokojivě sleduje trendy, nicméně při některých extrémních hodnotách sledované veličiny predikuje špatně. Tento jev lze vidět na následujících obrázcích.

V prvním případě, byť model náhlou extrémní změnu neodhadne, je následující predikce korigována a blíží se naměřené hodnotě. Naopak, existují extrémní hodnoty, ve kterých se predikce s časem nekoriguje.



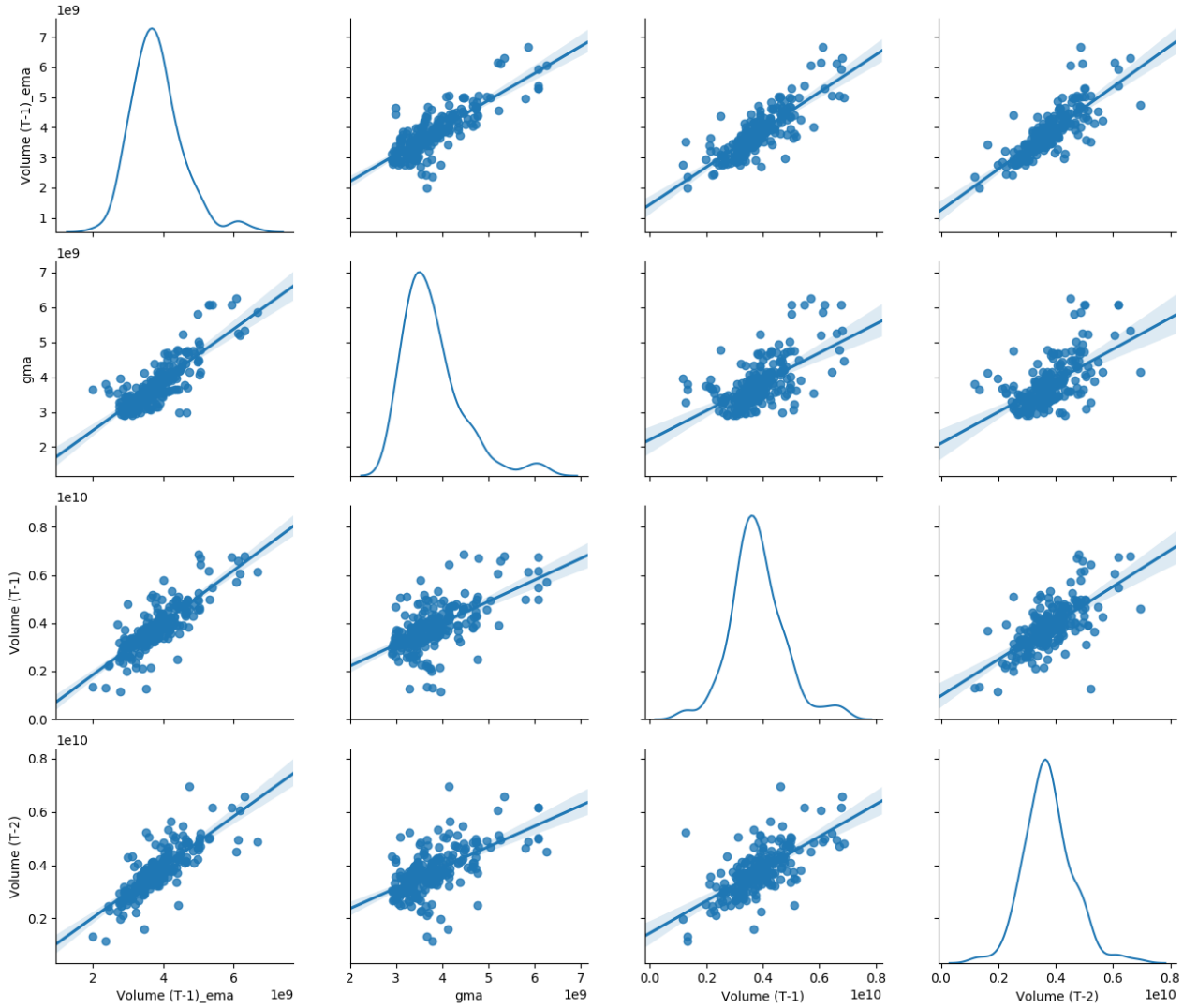
Obrázek: nahoře – část průběhu, ve kterém je model poměrně přesný a zdárně koriguje velké extrémy ve sledované veličině (Volume); dole – průběh ukazující části časové řady, ve kterých sledovaná veličina přechází mezi extrémy a model tento jev nezachytí.



Obrázek: Porovnání sledované veličiny (Volume) a predikce modelu (vlevo), resp. referenčního modelu (vpravo). Prokládaná křivka je polynom prvního řádu minimalizující MSE. Každý bod odpovídá jednomu vzorku testovací množiny.

Jiný pohled na přesnost je ukázán na předchozím obrázku. Zde lze vidět, že u velkých extrémů sledované veličiny model $f(h_d)$ podhodnocuje predikce, zatímco u nižších hodnot v_d nadhodnocuje. Lineární trend je zde zřejmý.

Naopak referenční model má větší rozptyl mezi sledovanou veličinou a predikcí při porovnání s modelem $f(h_d)$, což odpovídá výše zmíněným přesnostem R2. I tak ale referenční model dosahuje poměrně dobré přesnosti, což mimo jiné odpovídá relativně vysokému naučenému koeficientu pro složku v_d v navrženém modelu $f(h_d)$.



Obrázek: Scatter plot znázorňující vztah číselných příznaků mapy $\Phi(h_d)$. Popisky os: Volume (T-1)_ema je EMA(H_d), gma je MED(V_d), Volume (T-1) je v_d a Volum (T-2) je v_{d-1}.

Na obrázku výše je zobrazena závislost číselných příznaků mapy $\Phi(h_d)$. Zajímavá je nízká korelace mezi v_d a MED(V_d), byť dle předpokladu tyto veličiny měly být úzce spjaté. Skutečnost jsem pozoroval i při sestavování modelu. Přidáním složky MED(V_d) se významně zvýšila přesnost. Důvodem nejspíše bude vysoká nelineárnost této složky. Je možné, že přidáním jiných vysoce nelineárních složek by se zvýšila přesnost.

V. Co jsem vyzkoušel, ale nefungovalo to

Neuronové sítě se ukázaly jako nevhodný učící rámec. Zkoušel jsem různě hluboké verze postavené na fully-connected vrstvách s ReLU aktivacemi. Síť měla na vstupu příznakovou mapu $\Phi(h_d)$. Přesnost byla v některých oblastech časové řady srovnatelná s SVR, nicméně většinou byla spíše okolo referenčního modelu. Je možné, že jsem zanechal chybu v nastavení sítě, protože intuitivně by měly být sítě alespoň podobně přesné jako naučený afinní model.

Co se týče kernelů SVR, pracoval jsem i s nelineárními typy, nicméně zvolená příznaková mapa $\Phi(h_d)$ není vhodná pro Minkowské vzdálenosti. Proto pro všechny běžně používané nelineární kernely (např. rdf, polynomiální) by bylo potřeba vytvořit vhodnější příznakovou mapu, ve které by fungovala vzdálenost. Další možností je vytvořit vlastní kernel.

Měl jsem v plánu výslednou třídu modelů zapojit do ensamblu, zejména pomocí metody Adaboost, ale bohužel jsem si pro tento krok nevyhradil dostatek času. Letmý test ukázal, že Adaboost nad zvoleným modelem $f(h_d)$ se poměrně rychle přefituje. Nicméně kombinace regresorů do ensamblu je obecně dobrý způsob jak zvýšit přesnost.

Výběr složek modelu ukázal, že například sezonní závislosti nepřidávají na přesnosti. Jak kvartální, tak měsíční složka snižovala R2 přesnost modelu. Překvapivě, nepomáhalo přidání více kvantilů v $\mathrm{AQ}(V_d)$. Zvolené hodnoty $0{,}2$ a $0{,}8$ jako jediné zlepšily R2.

Často se osvědčí aplikovat tzv. dimensionality lift-up na složky lineárního modelu, čímž se zvýší kapacity modelu (ve smyslu VC-dimenze u klasifikačních problémů). Vyzkoušel jsem polynomiální zvýšení dimenze u hodnot $v_d, v_{d-1}, \ldots$, ale SVR se přefitovalo a R2 v některých oblastech časové řady spadlo do záporných hodnot.

VI. Závěr

Pro časovou řadu objemu obchodů v rámci indexu S&P 500 jsem vytvořil prediktivní model střední hodnoty sledované veličiny. V K-Fold vyhodnocení dosahuje model přesnosti 0,54 podle kritéria R2. Při Rolling Window vyhodnocení je hodnota R2 rovna 0,22. V obou případech je přesnost významně vyšší než zvolený referenční model.

Model je postaven na teoretických a empirických vlastnostech distribuce dat. Jedná se afinní kombinaci různých, vždy interpretovatelných složek, které jsem iterativně přidával do modelu vždy, když zvýšily přesnost. Koeficienty modelu byly kalibrovány metodou epslion-SVR, protože jiné metody poskytovaly horší přesnost. Empiricky bylo zjištěno, že model umí korigovat některé extrémní hodnoty na vstupu, nicméně je zde stále v tomto ohledu prostor pro vylepšení. To by například mohlo stavět na přidání více vysoce nelineárních složek, jakou byl výpočet mediánu.

Model lze v budoucnu vylepšit zohledněním dalších vstupních dat, např. změnou indexu či očekávaným významným eventem.

Ačkoliv je text česky, občas se schýlím k použití anglického termínu. Snad mi to jazykoví odborníci prominou.